DOCUMENT RESUME

ED 121 852 TM 005 284

AUTHOR Cone, Richard
TITLE Educational Implications of Criterion-Referenced Tests.
PUB DATE Apr 76
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, Calif., April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage
DESCRIPTORS *Behavioral Objectives; *Criterion Referenced Tests; Equivalency Tests; History; Program Evaluation; *Test Construction; *Testing Problems; Test Validity
IDENTIFIERS *Domain Specifications (Tests)

ABSTRACT

The educational implications of criterion-referenced tests are demonstrated. It is the hypothesis of the author that criterion-referenced tests have little educational impact unless carefully constructed around rigorous domain specifications. The paper details the process and problems of construction of a series of history tests presently being created under government contract. In addition, data is presented showing the effect of criterion-referenced tests on learners in classes being offered military personnel in Germany. It is anticipated that students will do singificantly better in classrooms where the teacher uses domain specifications. (Author/RC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

26.10

ED121852

Educational Implications of Criterion-Referenced Tests

A paper presented at the annual meeting of the
American Educational Research Association

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM005 284

April 23, 1976
San Francisco, California

Richard Cone
UCLA

During the past decade behavioral objectives have been touted as the technology which will rescue American schools from the mire of mediocrity. Commercial publishers, always anxious to help disseminate technology which will alter the world, have developed objectives, objectives-based tests, objectives-based assessment systems, etc. Schools have bought into systems which offer capabilities of diagnosing each student's strengths and weaknesses on a myriad of skill specific objectives. Such systems have, in many cases, presented a problem of "objective-overload" to the teacher. Fatigue is a very real factor to the teacher of reading with thirty to fifty objectives to be used diagnostically for thirty to thirty-five students. The thought of individually administering, marking, and recording more than 900 tests is enough to make most teachers stop and think, "Is it worth it?" The time taken for such massive testing can only be taken from one source, instructional time. The wisdom of diverting large chunks of time away from instruction is questionable.

In addition to being overly numerous, most objectives which are commercially available are overly narrow. The publishers have taken Gagné seriously and carried out task analysis ad absurdum. Thus, if a teacher wants to cut down on the number of objectives she is working with, she must choose between en route objectives of equal weight. If she attempts to cover all the objectives, she gets so caught up in details, that she loses sight of the large, important objectives and muddles on with no clear cut idea of where she and her students are going.

Meanwhile, tests of these overly specific objectives do not serve the needs or desires of the program evaluator. There is far more data than an evaluator can deal with. Limiting the number of objectives requires a decision about the objectives which might correlate best with the crucial skills which are the hoped for goals of a class or program.

Clearly, it is time to reconceptualize the implementation of objectives-based assessment in the classroom, yet a review of existing literature reveals a gross overemphasis on statistical validation of existing measures. Measures which are not serving the needs of the educational community may indeed be statistically "valid" but if they lack face validity (i.e., they will not, cannot be implimented properly under typical classroom conditions), to what avail are our efforts?

Jim Popham and a handful of other educators have begun a reassessment of the nature and role of objectives. The theory of these educators is that a "limited focus" strategy of testing

a few truly important objectives will enhance the ability to evaluate a program and will provide a manageable number of objectives for the classroom teacher (Baker, 1974; Millman, 1974; Popham, 1976). This has been a theoretical matter, there being no certainty that a few important objectives could be identified for a given subject or that tests could be developed for such objectives. This paper outlines a project that Popham and other U.C.L.A. faculty and students worked on which attempted to translate theory into practice.

In 1974 Jim Popham was contacted by educators working in the Predischarge Educational Program under the auspices of the United States Dependents Schools, European Area (USDESEA) about the creation of a series of tests which could be used to grant military personnel credits toward high school diplomas. It was specified that the tests would have to be referenced to skills and content taught in high school. Further, it was specified that the tests should allow for self-study or tutorial study as the primary purpose of the test was to allow men unable to attend regular classes an opportunity to gain equivalent credit. Lastly, it was hoped, but not specified, that the tests would provide focus and guidance for an extremely fluid teaching population. An agreement was reached to generate two tests - one in U.S. history, a second in third year English. The U.S. history test was to be an entire system of well-explicated criterion, sample items, teacher items, and the actual test items. The English test was to be only the criterion which PREP staff members would use in their generation of items. The author of this paper served as project supervisor for the history test.

Our first task was to determine what skills are required in most U.S. history courses. A review of texts and curricular guides led us to some important conclusions: 1) there is no agreed upon standard as to what constitutes a course in U.S. history, 2) most materials heavily emphasize content and the testing of content (i.e., memorization), and 3) the majority of history courses are tied to materials in that they begin on page one and continue to the end of the book or the end of the year, whichever comes first. Using this information, we were faced with the decision as to whether to generate a largely factual test which could be passed by high school students getting credit for a U.S. history course or to proceed with our own idealized version. At this point PREP began to get more interested in using our work as a guide to curriculum development efforts. They thus made the decision that passing the test would be equivalent to taking the course <u>as offered by PREP</u>.

Having thus rejected external criteria, we were forced to develop a set of criterion which could be applied to each of the objectives throughout the development of the test. One important criteria, instrumental in the exclusion of available sources, was a desire to generate a test which preserved a balance between higher and lower order cognitive processes. In

connection with this goal was an insistence on our part that each objective be sensitive to instruction. We wanted to avoid rewarding a student strictly on the basis of native intelligence or of some ability extraneous to instruction in history. To foster sensitivity to instruction, we decided to explicate each objective as fully as possible and practical. Our hope was that explicitly defining the full domain of the objective to be tested would force us to communicate to the user of the objective the precise skill to be utilized in item solution, the content parameters from which the items would be drawn, and the item format which would be used in all instances of testing the item.

Objectives, as conceptualized from their inception, had been intended to convey this information, but a single statement objective could do no more than vaguely outline a large domain or more clearly outline a very small domain. Popham had attempted to establish greater clarity in objectives through what he referred to as "amplified objectives." These objectives, while longer and more precise than objectives typically are, were not clear enough to avoid important ambiguities. A number of educators had proposed creating "domain specifications" in which a set format would be used to explicate a "domain." Hively had experimented with item shells in the '60's, but few other efforts had been made to generate working specifications (Hively, 1973). We were unable to make use of Hively's work or the thoughts of other educators in that generally they used math as the content area to illustrate domains. Math fits neatly into a conceptualization of clear and complete specification in that the contents and possible operations lend themselves to categorization. There is clearly a right way of doing things and a right answer. Such is not the case with history.

Our goal in generating domain specifications was to have guidelines for item construction which had conceptual rigor and behavioral specificity yet were brief and easy to use.* Most important we wanted to minimize ambiguities. Each of the domains was constructed so as to begin with a general description of the domain, which, like most objectives, tells what is presented to the examinee, what is done with that stimulus, and what is considered a correct response. Following the general description is a sample item which includes directions to the student and an item. This in turn is followed by a set of stimulus attributes. These are a set of rules necessary for generating all those elements the student must take into account prior to considering alternative answers. Response attributes generally conclude a set of domain specifications. Response attributes are a set

*In the Appendix there is an example of Domain Specifications prepared by the author and staff while working on the PREP history project.

of rules necessary for generating those factors the student must take into account while considering which of the answers is correct. These rules tell item writers how to construct both the correct answers and the distractors. In many cases, domains include a supplement which clarifies the content parameters.

While we had agreed to prepare the history test, we had no desire to tell the PREP history teachers what were the essential skills and content areas in history. The first months of the project were spent corresponding with a curriculum-testing committee made up of PREP teachers working in Germany. Our first major problem was communication, as it took typically three weeks to receive a response to an inquiry. This communication problem was more than a matter of physical difference, however. Semantics, lack of agreement, professional pride, interpersonal problems, and differing levels of expertise hampered the development of the test specifications.

One of our earliest problems was created by the inability of history teachers in the PREP program to narrow down the number of objectives that could be categorized as terminal. Our first communication from Germany brought us a multi-page document. The document split history into eleven content areas (e.g., basic facts, map skills, charts and graphs, etc.). Each of these was sub-divided into from two to six skill levels (e.g., knowledge, application, analysis, etc.). Finally, each skill level was accompanied by from one to three statements that roughly approximated behavioral objectives. Thus, for instance, a given content area, "Historical Documents," might be sub-divided into a number of skill levels, e.g., knowledge, comprehension, application, synthesis, and evaluation. These might be further sub-divided into statements such as "identifies purpose of the document," or "describes frame of reference."

From that first multi-page document a number of things became apparent. Perhaps most apparent was the fact that the PREP committee was considering too wide an array of objectives for purposes of either accurate measurability in a single three hour test or for optimal teaching focus. While we possessed information on the appropriate number of items for the determination of mastery on the part of an examinee (Millman, 1972), we had no empirical evidence as to the number of objectives that a teacher could realistically focus in on in the course of a year. Prior to starting the project, we knew both intuitively and from experience that teachers cannot use 35, 50, or 200 objectives according to the full mastery model without devoting a disproportionate amount of time to testing and bookkeeping. Our estimates of what might be realistic ranged from six to fifteen, but we decided to use ten for our magic number as we felt the objectives should be important enough to merit approximately

a month's instructional time. Please note that there was nothing empirical about this. No empirical information exists according to our knowledge on the effect of different numbers of objectives on the teachability of those objectives. We were not irrevocably committed to ten objectives (we eventually ended with eight) but merely used that as a rough target in helping the PREP committee focus in on a limited number of important objectives.

A second fact that emerged from our early communication with the PREP committee was that the history curriculum was loosely organized. Teachers were free to teach what they considered important. The first efforts by PREP to put objectives down in black and white had obviously led to a great deal of creative thinking.

Despite these negative attributes, this initial document provided some valuable information. It delineated eleven "areas," conveniently approximating our arbitrary decision to go with ten objectives. Most of these areas provided content guidelines such as "basic facts" or "historical documents." Skill levels in a given area were presented hierarchically so that the most appropriate skill could be easily evaluated. For instance, in an area called "Characteristics of a Given Period," levels of "knowledge, comprehension, and evaluation" were identified. From this point, it was comparatively easy to ask of the PREP teachers, "What do you really want the man to do with 'Characteristics of a Given period'?" Upon their response of "be familiar with basic periods in American history," it became clear that they were primarily interested in the knowledge level for that area. The idea was thus to transform their sometimes vague notions of their program objectives into statements which could be further transformed into measurable objectives. To this end, their initial taxonomy proved to be of some assistance.

The taxonomy also proved to be of assistance in helping us in the maintenance of a balance between higher and lower order cognitive skills. Some of the content areas clearly were most appropriately tested at the knowledge level (e.g., historical facts) while others were most appropriately tested at the analytical level (e.g., critical thinking). Content areas which could be tested at either a higher or lower level were then considered in terms of maintaining a balance on the overall test. Often we found that the skill which we felt was important enough to be included in one of our ten objectives combined more than one sub-skill. Thus, in the content area of map skills, we decided that the real objective about which PREP committee members were concerned was the ability to interpret historical maps. It was clear that to interpret maps a student would have to be competent at the knowledge level (e.g., identify map symbols), the synthesis level (e.g., generate information from symbols), and the analysis level (e.g., draw conclusions from the information generated). Within each of these levels it is possible to

conceive of a myriad of sub-levels, each sub-level with a myriad of item specific objectives.

As the above example illustrates, electing to test a more complex skill which required a variety of other skills caused us to sacrifice diagnostic ability for usability. We were making a decision to set a few well-defined targets for instruction and let professional teachers work on getting their students to the target.

We postponed facing the problem of commingled skills as long as possible by first working on the lowest order (i.e., easiest) domains. The area or domain identified to us by PREP as "familiarity with historic facts (people, places, events, and dates)" was narrowed down to "people" after much trans-oceanic communication about what it is fair to have a student memorize. "People," or "Significant Personalities in U.S. History" as it came to be called, asks the student to select from a set of personalities the one who could be credited with a real or hypothetical quotation presented in the item. Much of the construction of the domain specifications was mechanical in that we could easily specify the length and difficulty of the passage and what, in general, a quotation must contain in order for it to be linked to the speaker. Our problem was, "which personalities and what aspects of their personalities?" We asked the PREP committee to specify a set of personalities which would be encountered in PREP courses. Our test staff put together a list of personalities which they considered to be generally encountered in history classes. The names which were not on both lists were negotiated, categories of names added and deleted until consensus was arrived at. This list then was added to the domain as a supplement along with information about the selection process and the need for periodic revision of the list. The relationship between these personalities and the quote to be found in the item was made by relying on an external source (the Dictionary of American Biography). Rules for qualifying distractors were generated by relying on categories of mistaken identity (e.g., personality involved in the same event(s) as the correct personality but holding a different position or point of view).

Experiencing some success on this first attempt created within us the false belief that our task was going to be relatively simple. As we began to work on a number of others, we found the task immensely difficult and, on at least one occassion, impossible. True, two other domains of a lower order nature did go together with only a little more difficulty than the personalities domain. But as we attempted to capture the essence of critical thinking, or analyzing resource material, or even interpreting maps in a historical context, we faced a number of thorny problems. One of the problems which was particularly troublesome in generating a test in history was

the trivia problem. Anyone who has had a course in history knows the habit history teachers have of relying on minute facts for test items. We wanted to control for this by making explicit provisions for the quantity and quality of factual recall necessary for item solution. In some cases, where we felt we could get a fix on a student's ability to analyze or hypothesize historical concepts without the inclusion of historical facts, we generated our own hypothetical history. In other domains, for instance, "Personalities," we delineated an exhaustive list of basic facts (names) and guided item writers to a single source for specific information about those facts. In still other domains, we succeeded only in warning item writers not to deal in trivia and giving examples of the level of factual information we believed to be fair. Of the three approaches, PREP committee members were most shocked with the use of hypothetical history.

The use of hypothetical history first arose in our map domain. We had been asked to generate a domain which called on the student to evaluate historical maps and to draw conclusions based on his observations. One approach after another failed because of a single problem: the student who was knowledgeable about U.S. history need not even look at the map to see the relationship of geography to the western movement or the role of northern industries in the Civil War. We found after a number of weeks of individual and group effort that the problem remained. We also found that very often in history, geography did not play the role it should have. Every possible obstacle stood in the path of the western movement, yet within a relatively short period man had conquered every one. We were on the verge of testing map skills as a lower order activity when we were struck with the idea of generating our own history on sets of six maps, maps which the student would have to study in relationship to one another if he were to determine historical forces, movements, and their causes. This solution made it possible to generate multiple items per set (five) and guaranteed that mastery was due to skill in analyzing maps rather than to factual knowledge about U.S. history. Having decided this, specifying the types of information to be presented on the maps was a relatively simple matter as was the specification of the qualities of distractors.

Our brainchild met with immediate and unqualified rejection on the part of the PREP committee. We threw up our hands in desperation, told them of the only available option we could see, and asked them to consider alternatives. After a brief (but we understand stormy) period, they concluded that we had been right. The domain was completed with hypothetical history.

In other domains we could only justify them as appropriate for history students by the inclusion of historical facts. One such domain was entitled "Thinking Critically." We identified six logical fallacies which are common in the social sciences. Obviously we could not use advertisements or everyday experiences for our test items. The only logical course was to place the items in a historical context. This raised the issue of the necessity of thinking critically about the use of historical facts. It was decided to generate items in which the facts could be right or wrong and the logic fallacious or non-fallacious. Our alternatives were thus a fixed set of "factually accurate and non-fallacious, factually accurate but fallacious, factually inaccurate but non-fallacious, and factually inaccurate and fallacious." We then set about trying to control for the classes of facts which might be included in an item. We were, I believe, reasonably successful at specifying classes of factual error (e.g., placing an event more than 25 years out of context), but could only "control" non-erroring facts by warning and example (e.g., "an event of sufficient magnitude that a reasonably well-informed student. . ."). Had this been our first domain, it would have been a blow to our collective ego to realize there is a practical limit to specificity. Coming as it did late in the project, we had been well prepared for humbling experiences.

The completed "Thinking Critically" domain is a good example of a completed domain which represents important, real life skills. The problem, of course, is that real life skills are commingled skills. If a student fails to show mastery on the "Thinking Critically" domain, we are not sure whether it is because he does not know the fallacies, does not know his history facts, or does not read carefully. In return for the sacrificed diagnostic ability, we are putting into the hands of teachers explicit description of all aspects of the domain to be tested. The teacher can use professional judgement in preparing students for a test in order to insure prerequisite skills have been mastered. In the "Thinking Critically" domain, it would be up to the teacher to insure that all students could successfully recognize an ad hominem fallacy, a post hoc fallacy, etc., as well as have practice in examining statements for the various classes of factual error. The more important the domain, the more likely it is to subsume a number of skills and the more difficult it is to use the test as a diagnostic tool. This is not to say that the test is not a valuable aid to teachers. The specifications provide valuable assistance to the teacher who is attempting to help students achieve consequential goals.

The very visability of a few objectives and carefully delineated item generation rules prevents tests from being "cloud-referenced" (Popham, 1974). Objectives such as "The student will infer the point of view of a political cartoonist from a historical cartoon" are easy to state and the items easy

to generate. But when item generation rules are required, a great deal of hard thinking must be done in order to clarify such things as "infer," "point of view," and even "historical cartoon." This hard thinking required for the production of domain specifications is long overdue in testing. In history, traditionally we have tested at two levels - memorization and native intelligence. History does not lend itself to objective types of items except at the most basic (and base) level. Much work has been done in the new social studies, but the work has been curricular. Testing in social studies has not changed since the time of Oliver Cromwell.

In preparing the history test for PREP, the demand for specifications made it impossible for us to produce test items for an objective which all history teachers believe to be vital: the analysis of cause and effect. The PREP committee wanted cause and effect to be specified as a higher order cognitive objective. They were not interested in merely testing students' abilities to recognize commonly accepted causes and/or effects related to a given event. They wished to have the student do some analysis and select "the most logical cause or effect" of an event. We realized that if we did not test well-known (i.e., memorized) causes and effects, we had to explicate the qualities of causality so that we could truly say that a single cause or effect was more "logical" than another. We should have realized immediately from the lack of agreement among historians that cause and effect are open totally to interpretation, and that what appears to be logical to one appears absurd to another. We did not, however, readily recognize that fact because we were, at that juncture, inamored with a belief in our power to specify anything and everything. We referred to historians; we referred to logicians; we referred to philosophers. Each led us to the same conclusion: causality (with reference to historical events) exists in the eye of the beholder. Yet, we remained dissuaded. We tried to operationally control "the most logical" cause or effect by specifying that the distractors had to be clearly inferior to the correct alternative. We played with distractor definitions such as, "lacking contiguity in time or space," only to find that it was not illogical to think of the abolition of slavery as an effect of the California gold rush. We considered saying that distractors might be logical only through an involved chain of causality while correct alternatives would be immediately related. Specification of "immediate" posed a problem which was not dissimilar to the larger issue we were facing. Additionally, trying to specify the correct response by the creation of a large gap between that response and other responses made the items of questionable technical merit, i.e., more a measurement of native intelligence than of any sort of teachable skill.

In the end, we had to face the fact that we were beaten. Countless items constructed during the many attempts to specify were, in addition to being highly heterogeneous, either testing

knowledge of historical trivia or were testing some sort of native ability to determine a normed response (i.e., "seven out of ten historians agree that X was the cause of Y"). PREP was given the choice to either (1) specify those causes and effects they believed important enough to be tested as a lower order skill, or (2) drop the objective. They took the latter course of action.

As the other domains neared completion, they were subjected to a long and arduous process of in-house review. Initial review was done by a reader to locate specifications which did not communicate what they had been intended to communicate. Following this surface review, an attempt was made to use the specifications for item generation. Items generated by someone other than the writer of the specifications were checked against items prepared by the writer. Lack of item homogeneity, either in format or in difficulty, was interpreted as a failure of the specifications and further work was undertaken. Our last in-house review was another item generation, this time by a history specialist using the specifications minus the sample item. If the specifications led to item homogeneity without the sample item, we were reasonably well assured that they were working properly.

Our external review consisted of a review by a person with degrees in philosophy and law who read the specifications for such things as inconsistencies and ambiguities. What we had believed in-house to be crystal clear often was far from it, but all of us had been too close to the development of the specifications to see the problems. Following this review and extensive revision, each domain was carefully reviewed by PREP. The PREP committee members looked at the domain specifications to insure that they were consistent with the wishes of the committee in terms of content, skill, and format. Committee members were asked to try to generate items from the domains to determine the usability of the specifications by people at the classroom level. Following PREP's review was a final revision and review.

All of this occurred prior to any formal item generation and underscores the critical difference between CRT's with domain specifications and CRT's without. The above process is long, taxing, and costly - too costly to be done with a large number of objectives in a subject. On the other hand, the process of constructing rigorous specifications goes a long way toward insuring item homogeneity, clarifing what is really being tested in an item, and providing an excellent profile to the teacher of what the targeted skill really consists of.

The issue of item homogeneity plagued us somewhat in the course of test development. We faced this issue early when deciding to go with a few important objectives that subsumed

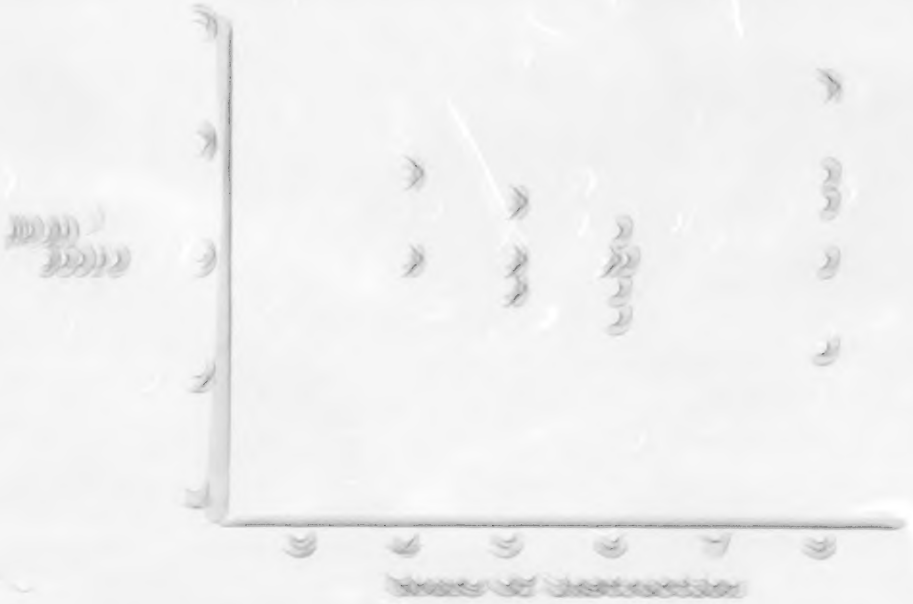
others. We continued to face it down to the generation of the

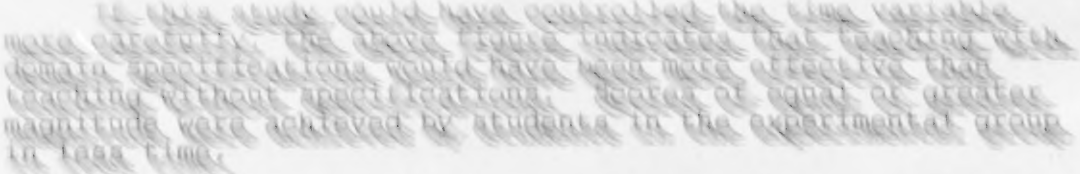
A few problems remain in some of the domains, problems largely due to an inability to specify clearly and explicitly enough to provide adequate guidance to an item writer. There are areas where a bit more time, and a bit more money, might make some difference. There are other areas where greater explication might not provide practical assistance to the item writer. There are areas now explicated where such explication is probably not necessary. There has been no research into the amount of explication that item writers need to produce homogeneous items. There is no proof that domain specifications are in any way superior to objectives. Philosophically, we held it to be an article of faith during the PREP project. It simply made sense that greater control over the items would lead to better, more

homogeneous items. Pragmatically, we became even more convinced

The results were somewhat disappointing. We had anticipated that even with the small number of items given as the posttest, students would be significantly better in classes where the teacher had access to the domain specifications. Such was not the case when results are tabulated as they should be, with the classroom as the unit of analysis. Both the experimental and control groups scored a mean of 2.7 on the pretest. The experimental classes scored better on the posttest with a mean of 3.33 while the control classes scored a mean of 3.05. This variance is not significant with an n of 15 (one of the teachers in the control group withdrew from the experiment).

One of the interesting aspects of the results can best





The coordinator of the experiment in Germany explained the discrepancy between our requested ten hours and the resulting four to eight hour effort as a result of negative attitudes on the part of teachers in both groups. Teachers apparently were uncomfortable teaching toward competencies on a "modular" basis in a course which has historically been taught chronologically. As the primary focus of the FREP project was the creation of an equivalency exam, this complaint does not undermine the work done. It does, however, pose problems for those who would use a mastery model in guiding the teaching of history. "The Civil War" can hardly be called a competency. One approach the history committee in Germany is considering, and one which has much appeal, is a cyclical competency approach. Thus, early

in the course students would work on mastering the historical terms, the personalities, the documents, etc., of the colonial period. Having mastered these, they would go on to the terms, personalities, documents, etc. of the constitutional period. During each period, the testing of a given competency would remain consistent with a single domain specification.

The study raises a number of interesting questions about the use of domain specifications while leaving open to question the value of such specifications. Is teaching with domain specifications more effective (over time) than teaching with an objective only? Does a teacher with domain specifications get bored with highly directed teaching and move on to other material for his or her own sake, without regard to the students' attainment of mastery? Would specifications for more abstract competencies (e.g., Critically Thinking in Historical Contexts) be more or less successful than the domain utilized in this experiment? These questions and a host of others may well be answered as the PREP program continues. At the same time, it is essential that more controlled experiments of this nature be implemented in order that the value of objectives in general and domain specifications in particular may be empirically verified.

The intent of this paper has been to outline what, to me, is a necessary next step in the objectives-CRT movement. I believe that it is unwise to hold teachers accountable for student progress on each of a multitude of objectives. It is possible to evaluate their performance through the use of a few truly consequential goals. I believe that it may be counter-productive to the education of children to focus our educational efforts constantly at the level of many narrow objectives rather than maintaining a focus on those skills that we see as being of real consequence. Even if one is convinced (which I am not) that the sum of the parts equals the whole, one must believe that the whole will be assembled by a classroom teacher harassed by the mechanics of an objectives-based assessment system. I favor providing a clear focus to a few objectives as a method of maintaining teacher accountability and, at least partially, restoring professional integrity. Teachers now possess the diagnostic tools to use informally or formally. What they lack are tests and objectives that will provide a focus on skills that are worthwhile in and of themselves.

Creation of such tests by methods similar to those outlined in this paper almost guarantees that limited focus objectives are both possible to instruct to and to measure. The generation of domain specifications forces subject matter specialists to think long and hard about what are the most important goals of a program as well as whether such goals are a realistic end product of classroom instruction. This thinking about importance and "teachability" is long overdue.

APPENDIX: SAMPLE DOMAIN SPECIFICATION

Application of Historical Terms

General Description

From a description of a historical event, the student will be able to select from a set of four alternatives the term that is most descriptive of that situation.

Sample Item

Directions to the Student: Below is a brief description of a situation in American history. Following the description is a list of four terms. Select the term that most clearly applies to the event. Mark the letter of the term you have selected on your answer sheet.

1. The American government controlling the Phillipines against the wishes of the Filipinos is an example of:

   a. imperialism
   b. isolationism
   c. nationalism
   d. federalism

Stimulus Attributes

1. The description must be of a historical event taken from any widely used U.S. history text. The event must exemplify one of the terms in the appendix according to the historically applicable definition of that term in Webster's Dictionary.
2. The description must end in the words ". . . is an example of:"
3. The description will be no longer than 50 words and represent an eighth grade reading level.

Response Attributes

1. Each of the four alternatives will be a single word or phrase (e.g., 'capitalism,' 'manifest destiny').
2. Each of the three distractors will be one of the following:

   a. A term drawn from the appendix and related economically or politically to the event described in the stimulus.

b. A term which is a likely cause or effect of the event described in the item (e.g., a "revolution" may be the effect of a "tariff").

c. A term that describes a step in a process that comes after or before the step actually described in the item ("impeachment" - "conviction').

d. A term which involves processes similar to those implied by the correct term but which is not associated with the group or goods delineated in the item (e.g., both "monopoly" and "socialism" include the concept of one group controlling an industry or business).

e. A term which involves the same group or goods as does the correct term but is not an example of the phenomenon described by the item (e.g., "tariff" and "embargo" both can be applied to the same commodities).

f. A term that has a definition opposite to that of the correct term (e.g., "depression" - "inflation").

3. The correct response will be that term whose definition (as it appears in Webster's Dictionary) is consistent with the description of the event [e.g., "imperialism" (the policy or practice of extending the power of one nation over the political or economic life of other areas) - "The American government controlled the Phillipines against the wishes of the Filipinos."].

4. Care should be taken to insure that only one alternative can be interpreted as being consistent with the description of the historical situation. It is recommended that more than one individual check to insure that no distractor could, in fact, be interpreted as a correct response.

Domain Supplement

Terms

Terms herein included were selected by the USDESEA PREP U.S. History committee (1975) based upon their use during more than a single period in American history. The meaning of the terms will remain constant while the manifestation of the term in history may vary.

boycott
capitalism
civil rights
communism
compromise

democracy
depression
discrimination
embargo
federalism

immigration
impeachment
imperialism
industrialization
integration

intervention
isolationism
labor union
laissez-faire
manifest destiny

monopoly
nationalism
political party
prejudice
reform

revolution
socialism
state's rights
strike
tariff
totalitarianism

Bibliography

Baker, Eva. "Beyond Objectives: Domain-Referenced Tests for Evaluation and Instructional Improvement," Educational Technology, June, 1974, 14:10-16.

Hively, Wells, et al. Domain-Referenced Curriculum Evaluation: A Technical Handbook and a Case Study from the Minnemast Project, CSE Monograph Series in Evaluation, No. 1, Center for the Study of Evaluation, University of California, Los Angeles, 1973.

Millman, Jason. "Program Assessment, Criterion-Referenced Tests, and Things Like That; Domain-Referenced Tests," Educational Horizons, Summer, 1974, 52:188-92.

Millman, Jason. "Determining Test Length, Passing Scores and Test Length for Objectives-Based Tests," Instructional Objectives Exchange, Los Angeles, 1972.

Morgan, Frank. "A New Alternative in Criterion-Referenced Test Development," paper presented at the 54th annual conference of the California Educational Research Association, San Diego, November 12, 1975.

Popham, W. James. "Procedural Guidelines - Developing IOX Objectives-Based Tests," Technical Paper No. 8, Instructional Objectives Exchange, August, 1972.

Popham, W. James. "Approaching Peril: Cloud-Referenced Tests," Phi Delta Kappan, May, 1974, 55:614-15.

Popham, W. James. "Merits of Limited Focus Criterion-Referenced Measurement Strategies," a symposium presentation at the annual meeting of the American Educational Research Association, San Francisco, April 19-23, 1976.